2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Cool Blog Identification using Topic-based Models

Kritsada Sriphaew, Hiroya Takamura and Manabu Okumura
Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta Midori-ku Yokohama 226-8503 JAPAN
kong@lr.pi.titech.ac.jp, takamura@pi.titech.ac.jp, oku@pi.titech.ac.jp

Abstract

Among a huge number of blogs on the internet, only some of them are considered to have great contents and worth to be explored. We call such kind of blogs cool blogs and attempt to identify them. To solve the cool blog identification problem, we consider three assumptions on cool blogs: (1) cool blogs tend to have definite topics, (2) cool blogs tend to have sufficient amount of blog entries, and (3) cool blogs tend to have certain levels of topic consistency among their blog entries. Corresponding to these assumptions, we extract a mixture of topic probabilities using a topic model, exploit the number of blog entries of each blog, and calculate the topic consistency among blog entries using distance functions over topic probabilities, respectively. We show the benefits of the proposed assumptions through these features. A feature unification model is also presented to achieve highest effectiveness. The experimental results on Japanese blog data show that we can improve the classification results by applying proposed assumptions.

1. Introduction

A blog is generally known as personal web pages usually maintained by an individual. Each blog consists of a sequence of blog entries in reverse chronological order. A rapid increasing number of blogs has leaded to not only huge amount of informative contents, but also abundance of uninteresting contents from the readers' viewpoint. Finding the cool blogs with interesting contents can be a challenge. In this work, we consider the coolness of blogs in an aspect of their contents, but not their appearance. We can achieve a common agreement among various readers to identify cool blogs without considering individual interests. The following examples can give an idea of what kind of blog is cool.

A blog which always presents a review of sushi restaurants is considered to be cooler by majority of readers than the other blogs which sometimes write about sushi restaurants that were visited by bloggers. Another example is that the blogs whose topics are easy to understand, such as new books, novel products or modern fashions, are considered to be cooler than blogs on some excessively technical topics, such as academic or company discussions. Although the interestingness on the content of cool blogs is subjective to the readers, the blogs with easy-to-understand topics are likely to draw attention from the readers because of belief on blogger's expertise and consistency of the blog's topic.

To identify cool blogs, we can simply represent the problem as a binary classification problem. Several classification techniques can be applied, but the main problem is that the suitable features for this task are unknown. Therefore, we start from an empirical investigation on cool blogs and introduce the assumptions on cool blogs. Following the assumptions, we employ topic-based models to extract features. The benefits of the proposed assumptions for cool blog identification are shown by the experiments.

2. Related Work

The interestingness of web contents has been studied in information retrieval [1, 7]. However, their target is reader-dependent interestingness of web sites regarding reader's profle, while ours is reader-independent coolness of blogs. A similar task is the credibility assessment of web data that tries to capture the trustworthiness and expertise of the web sites [3, 9]. Although the credibility differs from the coolness, the indicators for credibility can be used as clues for identifying cool blogs.

Rubin et al. proposed indicators based on four profile factors for the credibility of blogs [8]; 1) blogger's expertise and offline identity disclosure, 2) blogger's trustworthiness and value system, 3) information quality, 4) appeals and triggers of a personal nature. However, some of the indicators cannot easily be used for the actual application. For example, indicator such as honesty is hard to capture with current NLP techniques. Some indicators require blogger's personal data, such as names and addresses, which are often concealed. Weerkamp et al. [10] reported that some indicators can improve their application on blog post retrieval.

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.401

Mishne [6] regarded linkage patterns to identify bloggers with a similar interest. However, there is an argument over the effectiveness of link information [4]. Many kinds of features can be applied to cool blog classification; text-based (e.g., bag-of-words, n-grams), linguistic (e.g., part-of-speech), link-based (e.g., pagerank). Since it is impractical to try all possible combinations of the features, only features reflecting the proposed assumptions on cool blogs are investigated. In this work, we will set the baseline to be the classifier with bag-of-words features.

3. Assumptions for Cool Blog Identification

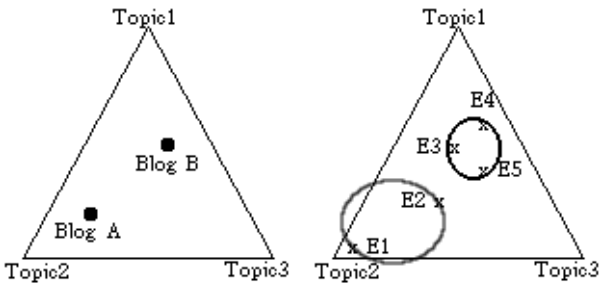
On the basis of empirical investigation, we propose three assumptions to identify cool blogs as follows.

1. *Cool blogs tend to have definite topics.* The readers judge the coolness of blogs from their perceived information which is obtained easily if the readers have experience on or understand the contents. The topics that are easy to understand by majority of readers tend to be common topics. For instance, topics related to comic, food, restaurant, sport, travel are comprehensible in common by majority of readers more than diary blogs, company discussion blogs or technical specific topics (e.g., NLP research blogs). We do not claim that this assumption is always true but it frequently occurs when we investigate the common intuition of coolness judged by different readers.

2. *Cool blogs tend to have sufficient amount of blog entries.* We make an assumption that the number of blog entries of each blog is one indicator to identify the cool blogs. Cool blogs tend to have some amount of blog entries rather than a few number of blog entries.

3. *Cool blogs tend to have certain levels of topic consistency among their blog entries.* Besides the topics covered by cool blogs as in the first assumption, cool blogs tend to focus on a solid interest rather than fluctuating topics among their blog entries. The consistency of topics in a blog can be observed by the changes of topics between its blog entries. Cool blogs tend to have a few changes on the topics of their blog entries.

We illustrate these assumptions using topic simplex diagram [2] as shown in Figure 1. In these topic simplexes, we assume that there are three topics where each corner of the simplex corresponds to the probability one for a particular topic and zero for the other topics. A blog can be represented as a point in topic simplex which means that it is a mixture of topic probabilities over all topics in the simplex. Assume that we have two blogs, i.e., A and B, as shown in Figure 1(a), A relates to $Topic2$ more than $Topic1$ and $Topic3$, while B relates to $Topic1$ and $Topic3$ more than $Topic2$. From the first assumption, we assume that there is a relationship between cool blogs and the topics. The cool blogs tend to be related to some definite topics. If we assume that $Topic$ 1, 2 and 3 are related to travel, diary and sport, respectively. It is possible that cool blogs will contain the mixture of topics with high probability on $Topic1$ and/or $Topic3$ more than $Topic2$.



(a) over Blogs (b) over Blog Entries

Figure 1. Topic Simplex.

Since a blog is composed from several blog entries, we can represent each blog entry as a point in the topic simplex shown in Figure 1(b). We assume that E_1 and E_2 are blog entries of a blog A, while E_3, E_4 and E_5 are blog entries of a blog B. Our second assumption states that cool blogs tend to have sufficient number of blog entries. This comes from the notion that sufficient amount of blog entries can convey sufficient contents used for identifying the cool blogs.

In the third assumption, it is assumed that the topics of cool blogs tend to be consistent among their blog entries. Therefore, we consider the consistency of topics in each blog. The topic consistency of a blog can be sketched as a circle that covers all its blog entries in the topic simplex as shown in Figure 1(b). Regarding to the assumption, cool blogs are likely to have smaller circles rather than the larger ones. Then, blog B tends to be cooler than blog A.

4. Implemented Models

We implement the models to extract feature sets regarding each assumption. For the first assumption, we explore a topic model to estimate the mixture of topics for each blogs. The number of blog entries is easily retrieved for the second assumption. Finally, to apply the third assumption, several distance measures are used for computing topic consistency among the blog entries of each blog.

4.1. Topic Model

Bag-of-words or vector space model has been widely used in the area of NLP and information retrieval for mod-

eling texts. However it is interesting to view contents as a probability distribution over topics. In statistical language processing, it is possible to treat a blog as a probabilistic mixture of topics where each topic is a probability distribution over words in a blog collection [5]. There are several choices of topic models that have received attention recently, but we will focus on one generative probabilistic model - Latent Dirichlet Allocation (LDA) [2]. Its idea is to represent the documents as random mixtures over latent topics and each topic is characterized by a distribution over words. In this work, we apply LDA model for estimating the topics from our large set of unlabeled blogs and determining the mixture of topics for each labeled blog. This representation for encoding a blog as a mixture of topics instead of conventional bag-of-words is applied, since the first assumptions believes that the indicators of coolness can be better achieved through the topic level than the word level.

For LDA, all parameters are automatically estimated by the model except the number of topics that we must specify. Although each mixture of topics that is inferred by LDA is not exactly a point (but a smoothing distribution) in the topic simplex as in Figure 1, we can assume its mixture as a representation of each blog where the summation of probability over all topics in each blog is equal to one.

4.2. Topic Consistency

This section presents several measures to model the topic consistency among blog entries of each blog. We explore three measures to reflect the topic consistency based on distance functions: Euclidean (EU) distance, Kullback-Leibler (KL) distance, and Jensen-Shanon (JS) distance. In the first step, we turn to the question of how to measure the distance between two probability distributions. Assume that a blog B composes of a set of blog entries $\{E_1, E_2, ..., E_n\}$ where each blog entry E_i, $1 \leq i \leq n$, is represented by a mixture of T topics $\{e_{i1}, e_{i2}, ..., e_{i|T|}\}$, E_c is a centroid of blog B based on mixtures of topics over all blog entries of B where $E_c = \{e_{c1}, e_{c2}, ..., e_{c|T|}\}$; $e_{ct} = \frac{1}{n}\sum_{i=1}^{n} e_{it}$; $1 \leq t \leq |T|$. For any $E_i \in B$, we calculate the distance between a mixture of topics E_i and centroid E_c, denoted by $D(E_i||E_c)$, where the distance can be given by one of the following three choices of distance function:

$$\text{EU distance} : D_{EU}(E_i||E_c) = (\sum_{k}^{|T|}(e_{ik} - e_{ck})^2)^{1/2}$$

$$\text{KL distance} : D_{KL}(E_i||E_c) = \sum_{k}^{|T|} e_{ik} log\left(\frac{e_{ik}}{e_{ck}}\right)$$

$$\text{JS distance} : D_{JS}(E_i||E_c) = \frac{1}{2}\left(\sum_{k}^{|T|} e_{ik} log\left(\frac{e_{ik}}{e_{ik}+e_{ck}}\right) + \sum_{k}^{|T|} e_{ck} log\left(\frac{e_{ck}}{e_{ik}+e_{ck}}\right)\right)$$

In the second step to find the topic consistency of a blog, we simply take the mean of all distances between the mixture of topics and its centroid. Since KL distance is not symmetric, we calculate the distances for both $D(E_i||E_c)$ and $D(E_c||E_i)$, $\forall i, E_1 \in B$, and the mean to reflect topic consistency. The notations TC_{EU}, TC_{KL} and TC_{JS} are used to represent the topic consistency models which apply the mean of EU distance, KL distance, and JS distance, respectively.

5. Experiments

To study the effectiveness of the proposed assumptions, we conduct several experiments by exploiting different feature sets based on each assumption, and investigate the experimental results.

5.1. Experimental Settings

To the best of our knowledge, there is no benchmark corpus for cool blog classification. A blog search company, blogWatcher Inc., kindly provided large-scale data on Japanese blogs for this research. In the data, there are 838,341 blog entries on distinct 60,736 blogs. Their issued dates are varied from Feb 2005 to Dec 2007. Manual labeling of cool blogs was done by an engineer in that company, who is a native Japanese speaker, without any given assumptions. We do not even interview the annotator about the notion to label the cool blogs. This can avoid the bias of our assumptions to identify cool blogs. The annotator judged the coolness of each blog on a whole blog instead of an individual blog entry. The annotator selected some blogs in random order and categorized them into two sets, i.e. positive (cool) and negative (uncool) sets, with a balance of the number of blogs. In total, the corpus has 270 positive blogs (9,827 blog entries), 270 negative blogs (3,836 blog entries) and the rest are left as unlabeled. Although the dataset is a collection of Japanese blogs, we did not use any language-specific features in the task of classification. In other words, our contribution of this work can be applied for identifying cool blogs in any languages if the concept of cool blog is the same.

We used TinySVM[1], an implementation of SVM, as a classification model, with the optimized regularization parameter. Through all experiments, the linear kernel function, which is commonly used for text classification, was employed. All experiments were conducted in three-fold cross validation where two folds were used for training and the rest for testing, except one setting in Section 5.2.1. Each fold has 90 positive and 90 negative examples.

[1] http://chasen.org/∼taku/software/TinySVM/

Table 1. Baseline (bag-of-words features)

Trained by	Acc.	Prec.	Rec.	F
One fold	0.776	0.804	0.735	0.768
Two folds	0.787	0.819	0.741	0.778

5.2 Experimental Results

5.2.1 Baseline

We set the baseline as the case of bag-of-words features. All instances for learning the classifier are encoded by bag-of-words features of all vocabulary in the corpus. Two different settings are explored. In the first setting, one fold is used to train the classifier and test on either of the other two. In the second setting, we train the classifier by two folds of labeled data and test by the remaining fold. Table 1 shows the effectiveness of the baseline. This result presents the potential of this task if we want to construct the application to retrieve cool blogs because of its high precision. Although we can improve the effectiveness by increasing the size of training data, it is not possible to provide all labeled examples (reflecting by bag-of-words) that can cover all examples of cool blogs.

5.2.2 Mixture of topics

This experiment aims to identify cool blogs by applying a mixture of topics as a feature set for classification. The learning steps are as follows. First, the topic models are estimated from different sizes of unlabeled blogs with the specified numbers of topics using LDA. The topic models are estimated in the level of blog, where each instance is represented by a set of blog entries in a blog. Second, the mixtures of topics for both positive and negative blogs are determined by those generated topic models. For each blog, the mixture of topics is used as feature set for learning in a supervised manner. The results are shown in Table 2, where "#NB" is the number of unlabeled blogs used for topic model estimation and "Feature" indicates how we encode the feature set. For example, LDA_{20} is a set of mixture of topics features estimated from LDA with 20 topics.

Comparing with the result of baseline trained by two folds, we achieved higher accuracy and F-measure than the baseline in most cases. Moreover, the number of features in the case of mixture of topics (20, 50 and 100) is extremely smaller than the number of features in the baseline case (approximately 100,000 words in the vocabulary). To achieve better estimated topic model by a larger size of unlabeled blogs, we need to trade off with the computational time to estimate the topic model. In the latter experiments, we will explore only the cases of the topic model estimated from 1000 unlabeled blogs.

Table 2. Mixture of topics features

#NB	Feature	Acc.	Prec.	Rec.	F.
1000	LDA_{20}	0.750	0.726	0.800	0.761
	LDA_{50}	0.780	0.731	0.885	0.801
	LDA_{100}	0.767	0.727	0.859	0.787
3000	LDA_{20}	0.770	0.745	0.822	0.781
	LDA_{50}	0.791	0.778	0.815	0.796
	LDA_{100}	0.776	0.749	0.833	0.789
5000	LDA_{20}	0.761	0.746	0.793	0.769
	LDA_{50}	0.796	0.778	0.830	0.803
	LDA_{100}	0.791	0.787	0.796	0.792

Table 3. Combined feature sets: baseline/ LDA, the number of entries and the topic consistency

Feature	Acc.	Prec.	Rec.	F.
Baseline $+NE$	0.808	0.737	0.787	0.761
$LDA_{20} + NE$	0.807	0.837	0.767	0.800
$LDA_{50} + NE$	0.831	0.832	0.833	0.833
$LDA_{100} + NE$	0.817	0.802	0.841	0.821
$LDA_{20} + TC_{EU}$	0.800	0.759	0.878	0.814
$LDA_{20} + TC_{KL}$	0.783	0.746	0.859	0.798
$LDA_{20} + TC_{JS}$	0.789	0.746	0.874	0.805
$LDA_{50} + TC_{EU}$	0.817	0.776	0.893	0.830
$LDA_{50} + TC_{KL}$	0.802	0.755	0.893	0.818
$LDA_{50} + TC_{JS}$	0.807	0.772	0.874	0.820

5.2.3 Number of Entries

In this setting, the number of blog entries (NE) of each blog is used as a feature for cool blog classification. Since it is imprudent to apply this single feature alone for classification, we combine either bag-of-words (baseline) or mixture of topics (LDA) feature sets and NE feature which is weighted by λ. The results when applying this combined feature sets are presented in the first chuck of Table 3 where the "Feature" contains "$+NE$". Note that λ^{opt*} is the optimized weight to achieve highest effectiveness and can simply be obtained simply using training data. In this setting, λ^{opt*} is equal to 0.1.

Comparing with Table 2, the accuracy is improved approximately 6% in each case when the combined feature sets of LDA and NE are employed. Due to space limitation, we cannot show the other case of combined feature sets based on LDA that is estimated from 3000 or 5000 blogs, but there is small distinction between their effectiveness and the ones shown in the table. Note that the case of baseline achieves higher accuracy but lower precision, when NE is applied.

Table 4. Unification of feature sets: mixture of topics (LDA), number of blog entries (NE) and topic consistency (TC)

Feature	Acc.	Prec.	Rec.	F.
Baseline	0.787	0.819	0.741	0.778
$LDA_{20} + NE + TC_{EU}$	0.833	0.813	0.867	0.839
$LDA_{20} + NE + TC_{KL}$	0.811	0.839	0.774	0.805
$LDA_{20} + NE + TC_{JS}$	0.822	0.827	0.819	0.823
$LDA_{50} + NE + TC_{EU}$	0.833	0.831	0.841	0.836
$LDA_{50} + NE + TC_{KL}$	0.831	0.832	0.833	0.833
$LDA_{50} + NE + TC_{JS}$	**0.843**	**0.841**	0.848	**0.845**
$LDA_{100} + NE + TC_{EU}$	0.831	0.802	**0.878**	0.838
$LDA_{100} + NE + TC_{KL}$	0.817	0.802	0.841	0.821
$LDA_{100} + NE + TC_{JS}$	0.831	0.832	0.833	0.833

5.2.4 Topic Consistency

We explore the third assumption by applying each distance functions to encode the topic consistency (TC) among blog entries of each blog as a feature. In this experiment, the topic models for calculating topic consistency are estimated in the level of blog entries, where each instance is encoded as a blog entry. Thereafter, the mixtures of topics for both positive and negative data in the level of blog entries are determined by these topic models. It is expected that these topic models can capture fine-grained topics of blog entries, since our target is to detect the blogs that has solid interests among their entries. According to the distance functions presented in Section 4.2, different value of TC will be obtained. We combine LDA feature sets and TC feature using the same method as the NE case with λ^{opt*}.

In the second and third chunks of Table 3, TC_{EU}, TC_{KL} and TC_{JS} represent the topic consistency features which are calculated based on EU, KL, and JS distances, respectively. The results show that these features can improve the classification accuracy including the F-measure when comparing it with the results of applying LDA alone in Table 2. It can be seen that TC_{EU} performs better than TC_{JS} and TC_{KL}. Note that we cannot apply TC for the baseline since we do not have the information of topic probabilities in such case.

5.2.5 Unification of All Feature Sets

To study the effectiveness of applying all proposed assumptions for a task of cool blog identification, we combine all above feature sets; LDA, NE and TC, as follows

$$X_{new} = \{\lambda_1 X_1, \lambda_2 X_2, \lambda_3 X_3\},$$

where X_1 is LDA feature set, X_2 is NE feature, X_3 is TC feature. λ_1, λ_2 and λ_3 are constant weights for X_1, X_2 and X_3, respectively, which can be optimized by using training data. In practice, we set λ_1 to be one and optimize the values of λ_2 and λ_3. The results when applying several combinations of these feature sets are presented in Table 4. The result of baseline trained by two folds is also repeated in the table. The bold fonts in the table show the best performances that we can obtain when applying all proposed assumptions. There are significant improvements when the proposed features sets are used for cool blog identification.

6. Conclusions

The contribution of this work are as follows. We are the first to address the problem of cool blog identification using three assumptions observed from the property of cool blogs. The three assumptions are (1) cool blogs tend to have definite topics, (2) cool blogs tend to have sufficient amount of blog entries, and (3) cool blogs tend to have certain levels of topic consistency among their blog entries. We proved the benefit of assumptions by extracting feature sets that can reflect each assumption using topic-based models and investigating the results. To combine those extracted feature sets, we presented a feature unification model that takes all assumptions into account. The results convinced us that we can obtain effective results by applying the proposed assumptions for cool blog identification.

References

[1] D. Billsus and M. J. Pazzani. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27:313–331, 1997.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.

[3] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber. How do users evaluate the credibility of web sites?: a study with over 2,500 participants. In *Proc. of DUX-03*, pages 1–15. ACM, 2003.

[4] D. Hawking and N. Craswell. Overview of the TREC-2001 Web Track. In *Proc. of TREC-2001*, 2001.

[5] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of SIGIR-99*, pages 50–57. ACM, 1999.

[6] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, 2007.

[7] M. J. Pazzani, J. Muramatsu, and D. Billsus. Syskill & webert: Identifying interesting web sites. In *Proc. of NCAI-96*, volume 1, pages 54–61, 1996.

[8] V. L. Rubin and E. D. Liddy. Assessing credibility of weblogs. In *Proc. of AAAI-06: CAAW*, 2006.

[9] J. Stanford, E. R. Tauber, B. Fogg, and L. Marable. Experts vs. online consumers: A comparative credibility study of health and finance web sites. In *Consumer Web Watch Research Report*, 2002.

[10] W. Weerkamp and M. de Rijke. Credibility improves topical blog post retrieval. In *Proc. of ACL-08: HLT*, pages 923–931, 2008.